USCMS Facility Plans at FNAL

Jon Bakken

Computing TDR Tier 1 Services

Data archiving services

- Custodial share of raw (1st copy at CERN) + reconstruction
- All reprocessing datasets
- Full AOD dataset
- MonteCarlo from USCMS Tier 2s.

Large scale data analysis

- Sequential reconstruction and reprocessing
- Chaotic analysis activities
- Timely calibration feedback

Distribution of data to all Tier 1 and Tier 2 sites

User support

Site security

Prioritisation according to CMS policy

Accounting

Database Services

Computing TDR Tier 1 Requirements

CPU:

- 2.5 MSI2k,   2:1 ratio for scheduled/analysis activities

Data Disk:

- 1.2 PB, 85% reserved for analysis

Mass Storage:

- 2.8 PB, Max data loss ~ 10s of GB per PB
- 800 MB/s IO rate, written once, read many times

WAN:

- Structured: Incoming: 7.2 Gb/s Outgoing 3.5 Gb/s, (10.7 Gb)
  - Minimum requirements, expect more at FNAL
  - Assume controlled and highly structured environment

Models

Funding:

- Baseline
  - The minimal amount to safely perform task
- Reduced
  - Not quite enough to perform task, but not failure either
  - Hope to make up for deficit in later years
- Leadership
  - More than needed -- USCMS could dominate certain parts

Sizing:

- Minimal - USCMS is twice a nominal Tier-1 center. This is equivalent to 28% of total Tier-1 resources
- Fair Share - USCMS is 40% of collaboration, should have 40% of total Tier-1 resources

Tier 1 Sizing

Reduced Model

- 1/3 less CPUs for analysis activities
- 2/3 less data disk for RAW sample
  - More tape drives to handle deficit

Leadership Model

- 1/3 more CPU for analysis
- All the DST samples at FNAL Tier 1 site

Minimal & Reduced Fair Share produce ~equivalent numbers

- Default values USCMS is building the Tier 1 facility

- Funding for Tier-1 facility according to baseline is ~$8000k until 2008

Funding	Model	Characteristics	Cost
baseline choices	"minimal share" Tier-1	two nominal Tier-1 acc. to CM	$7600
	"fair share", reduced resources	40% share, less FEVT data	$7900k
leadership funding	"fair share", full resources	40% share including FEVT	$9270k
	"minimal share", leadership	two nominal Tier-1s + full DST	$8820k
reduced funding	"minimal share", reduced resources	two nominal Tier-1, less FEVT data	$6450k

CPU Projections

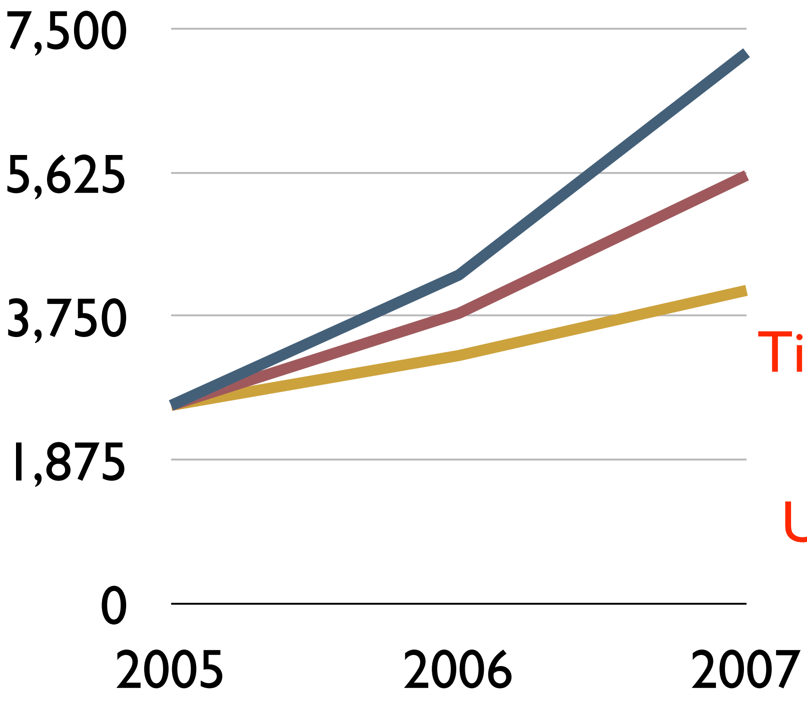
Lots of evidence that Moore's Law is breaking down for CPU scaling

- For example, roadmaps say we should be at 4 GHz now, cf 3.6 GHz
- We have measured Opteron 246 at ~2600 SI2K

Moore's Law CPU Scaling gives 4300 and 7200 SI2K for 06 and 07

A Pessimistic CPU Scaling gives 3250 and 4100 SI2K for 06 and 07

Compromise CPU Scaling gives 3800 and 5600 SI2K for 06 and 07



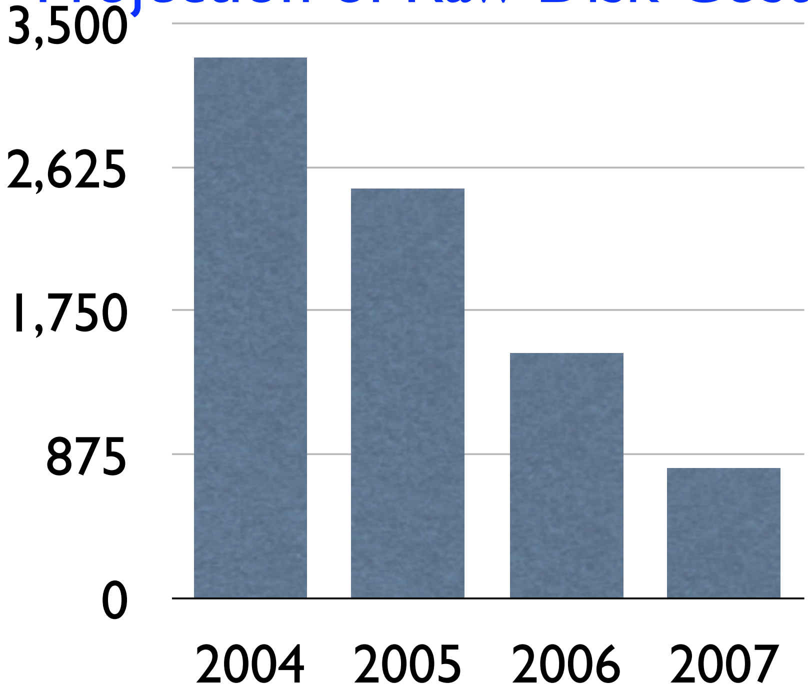
Projected CPU Performance

Tier 1 currently has ~175 worker nodes deployed with ~175 KSI2K

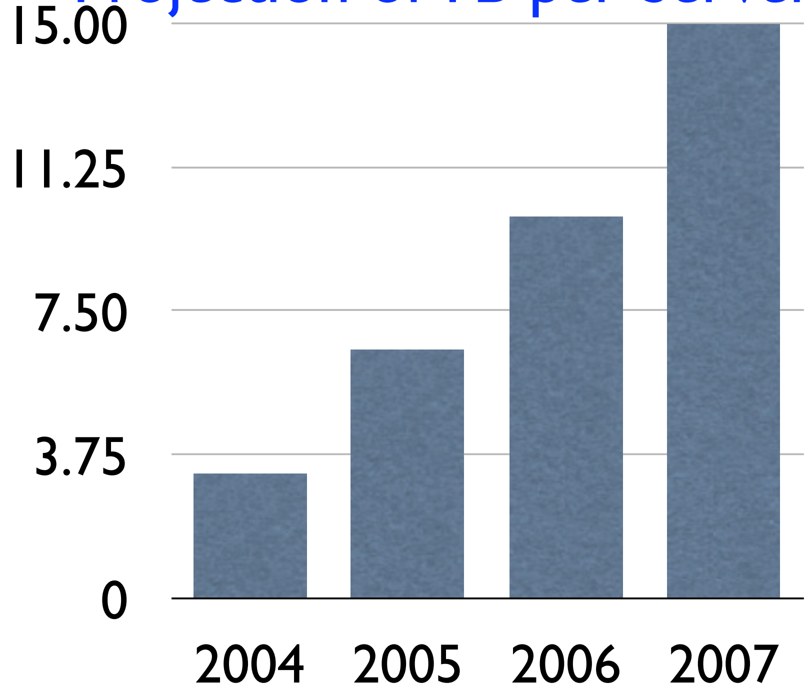
UAF has 53 nodes and 44 KSI2K

Data Disk Projections

Projection of Raw Disk Cost $/TB



Projection of TB per Server



Data disks have dropping costs per raw TB and the associated servers can serve more TB per unit in 06 and 07. Industry contacts indicate this disk scaling will continue

Tier 1 center now has ~80 TB of data disks currently deployed using dCache for posix & dccp onsite, gridftp offsite access

T1 Parameters

	Minimal Share	Fair Share Reduced
Worker CPU	4256 KSI2K 1000 nodes	4621 KSI2K 1078 nodes
Data Disks (TB)	1986	1546
Tape (TB/yr)	3224	3779
User Disk (TB)	20	20

Tape Projections

CMS currently has 1 9910 Powderhorn silo

- 5000 slots, currently 197 TB on tape
- This silo is part of the general FNAL tape infrastructure and connects to another 9310 shared silo that is shared by other experiments. Sharing has worked out well.

8 first-priority 9940B drives and 3 shared 9940B drives

- 200 GB tapes, 30 MB/s drives. See effective 20 MB/s rate
- Also 3 older 9940A, not used by CMS anymore

Expect to use archive 150 TB in 05 and 300 TB in 06.

- This corresponds to 750 tapes in 05 and 1500 tapes in 06, which safely fit inside our current 9310 Powderhorn.
- Also expect to store many PB of data during service challenges, but these fake data tapes will be quickly recycled.

Tape Projections

For CMS detector data in late 07, we need more robotic storage

- Need to buy in early 07 to have completely debugged

Expected drive is LTO, but have time before committing

- 800 GB cartridges for planning
- 120 MB/s, but use 50 MB/s effective rate for planning

Working closely with FNAL CCF Dept on robotics/tape choices.

All CMS data will go through a disk cache before being written to tape to help with rate adaption and file re-use by scientists.

- Do need to re-populate data that has been flushed from cache
- In reduced raw data disk cache models, need to have more drives to repopulate the cache more often, or be able to get data from another reliable source.

Networking

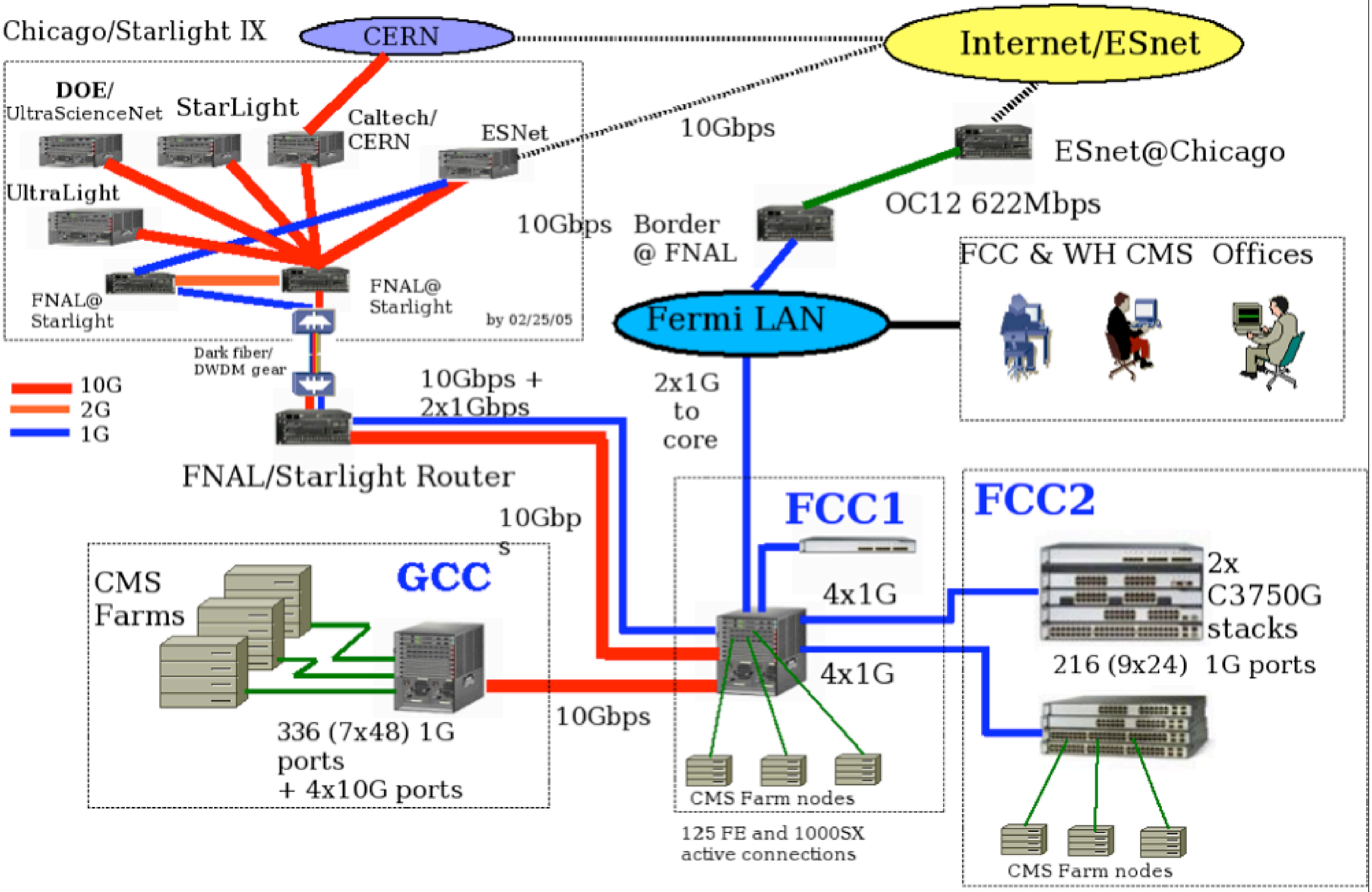
Expect structured incoming rate at 17 Gbits/sec,
and structured outgoing rate at 8 Gbits/sec
for a total of ~25 Gbits/sec for WAN traffic.
This is minimum for structured and organized data transfers.

We also expect LAN rates at 20 Gbits/sec, from the mass storage to worker nodes.

Tier 1 sites are required to accept data from CERN and distribute data to all Tier 1 and Tier 2 sites that request it.

CMS Service Challenges have achieved goal rates of 7 Gbits/sec incoming rates from CERN to FNAL.

CMS Network @ Fermilab (02/23/2005 AB)



CPU Plan

Worker Nodes

Year	Number of Nodes
05	280
06	320
07	478

Data Disk Plan

Disk Units

Year	Number of Units
05	8, now 5
06	45
06	43*

Fileservers

Year	Number of Fileservers
05	16, now 10
06	90
07	86*

Robotics Plan

Need to buy new robot, or have access to equivalent resources, in early 2007.

- Expect to write 3-4 PB/year

Analysis Servers Plan

Our experience has been that we need ~3-5% worker node count for analysis servers.

For example, in FY04, we estimated 20 servers and we bought 24 servers. These nodes were used for

- Development (DAG Cluster)
- Integration (CMS-ITB)
- General Purpose (gateways, databases, etc)

Our estimate is that we will be buying ~20 servers/year in FY05, FY06, FY07

Year	Analysis Servers
05	20
06	20
07	20

User Disk Plan

User Disk is a distributed global file system - we use IBRIX

- Expect 20 TB of User Disk at beginning of experiment
- It differs from data disk in that User Disk readily allows one to open and close files without penalty.

We have ~8 TB of User Disk, in 8 LSI disk arrays and 8 Dell file servers.

Year	LSI Arrays	Fileservers
05		
06	6	6
07	6	6